Target Selection and Deselection at the Berkeley Structural Genomics Center

45-letter short title: Target Selection and Deselection at the BSGC

Authors: John-Marc Chandonia[1], Sung-Hou Kim[1,2], and Steven E. Brenner[1,3]

Address for correspondence:
Steven E. Brenner
Department of Plant and Microbial Biology
461A Koshland Hall
University of California
Berkeley, CA 94720-3102
email: brenner@compbio.berkeley.edu
fax: (413) 280-7813

Affiliations:
1 - Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
2 - Department of Chemistry, University of California, Berkeley, CA 94720, USA
3 - Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

Keywords: BSGC, *Mycoplasma pneumoniae*, *Mycoplasma genitalium*

ABSTRACT

At the Berkeley Structural Genomics Center (BSGC), our goal is to obtain a near-complete structural complement of proteins in the minimal organisms *Mycoplasma genitalium* and *M. pneumoniae*, two closely related pathogens. Current targets for structure determination have been selected in six major stages, starting with those predicted to be most tractable to high throughput study and likely to yield new structural information. We report on the process used to select these proteins, as well as our target deselection procedure. Target deselection reduces experimental effort by eliminating targets similar to those recently solved by the structural biology community or other centers. We measure the impact of the 69 structures solved at the BSGC as of July 2004 on structure prediction coverage of the *M. pneumoniae* and *M. genitalium* proteomes. The number of *Mycoplasma* proteins for which the fold could first be reliably assigned based on structures solved at the BSGC (24 *M. pneumoniae* and 21 *M. genitalium*) is approximately 25% of the total resulting from work at all structural genomics centers and the worldwide structural biology community (94 *M. pneumoniae* and 86 *M. genitalium*) during the same period. As the number of structures contributed by the BSGC during that period is less than 1% of the total worldwide output, the benefits of a focused target selection strategy are apparent. If the structures of all current targets were solved, the percentage of *M. pneumoniae* proteins for which folds could be reliably assigned would increase from approximately 57% (391 of 687) at present to around 80% (550 of 687), and the percentage of the proteome that could be accurately modeled would increase from around 37% (254 of 687) to about 64% (438 of 687). In *M. genitalium*, the percentage of the proteome that could be structurally annotated based on structures of our remaining targets would rise from 72% (348 of 486) to around 76% (371 of 486), with the percentage of accurately modeled proteins would rise from 50% (243 of 486) to 58% (283 of 486).

Sequences and data on experimental progress on our targets are available in the public databases TargetDB and PEPCdb.

BACKGROUND

M. genitalium and *M. pneumoniae* were the first-sequenced members of the class *Mollicutes*, a group of wall-less prokaryotes distinguished by their small genome sizes; the latter characteristic has earned them the name "minimal organisms" [1,2]. Minimal organisms have been the subject of numerous experimental and computational genomic studies because of the possibility of identifying the minimal complement of genes necessary for life [3-5]. Because of their tractable size, organisms with minimal genomes have also been popular for structure and function prediction [2,6-13].

Structural genomics is an international effort to determine the three-dimensional shapes of all important biological macromolecules, with a primary focus on proteins. Most approaches involve coarse-grained sampling of protein families, aiming to provide one structure from each family, allowing folds of all family members to be recognized by homology [14]. Several strategies for selecting proteins as targets have been proposed, including selecting all proteins in single genome [15-17], selecting proteins that will allow a maximal number of sequences to be modeled at some level of reliability [18-21], or selecting proteins of biological interest such as those from important biochemical pathways [22] or those thought to be unique to a particular species (ORFans) [23]. Details of these target selection strategies have been reviewed extensively [14,24-29], and implications of future selection strategies are discussed elsewhere [30].

In the United States, the National Institutes of Health are supporting structural genomics projects at 9 pilot centers through the Protein Structure Initiative (PSI). Our work

is in the Berkeley Structural Genomics Center (BSGC), one of these 9 centers. The BSGC began in September 2000, and this is a report on progress to date. Our aim is to obtain a near-complete structural complement of the proteins in *M. pneumoniae* and *M. genitalium*. As *M. pneumoniae* proteins are largely a superset of the proteins found in *M. genitalium* [31], target selection is focused on the former proteome. Because of the relatively small size of these proteomes, it was expected that determining structures for most of the experimentally tractable proteins would be possible within the 5-year pilot period. Obtaining a near-complete structural complement of a single proteome would have the potential to enable new avenues of research that depended on this completeness. This would be analogous to the research into non-coding regions of DNA that has been enabled by the availability of complete genome sequences. Targets for the BSGC have been chosen in several stages: targets seen as "low hanging fruit" were attempted first, and later stages have targeted proteins predicted to be more experimentally difficult. Targets in later rounds were also chosen using more sophisticated bioinformatic analyses, such as domain prediction, which were not in place at the beginning of the project. Finally, target selection methods were refined somewhat, in response to early experience gained at our center and others. For example, in later target selection rounds more targets related to a single *Mycoplasma* protein were chosen to be experimentally studied in parallel.

One important aspect of target selection that was not fully appreciated until the project was underway was the need for target deselection. In the BSGC, we are only seeking to solve structures of proteins for which the structure can not be reliably predicted via bioinformatic methods. As new structures are constantly being solved by structural biology and structural genomics groups worldwide, it is necessary to frequently reexamine our target

list and remove targets for which the structures of similar or identical proteins have been solved elsewhere. We devised an automated procedure for identifying likely candidates for target deselection. These candidates are manually examined at weekly meetings to determine if they should in fact be stopped or whether the information that could be gained by finishing the structure is worth the effort. In this report, we examine the impact of target deselection and the reasons targets have been deselected.

Each round of target selection has led to successively more coverage of the *M. pneumoniae* proteome. In this report, we quantify the degree of coverage on two levels. First, we examine the percentage of the proteome that could accurately be modeled. This requires at least 30% sequence identity between the experimentally solved target and the *Mycoplasma* protein. Second, we estimate the percentage of the proteome for which the general fold can be predicted by homology with reasonable accuracy, whether or not there is sufficient confidence in the alignment accuracy to enable accurate structural modeling. The latter is described as "coarse" coverage of protein sequence space, and the former as "fine" coverage (see http://grants2.nih.gov/grants/guide/rfa-files/RFA-GM-05-001.html). Both of these percentages are calculated on a per-protein, where a protein is covered if any part can be structurally predicted, and per-residue, where we consider the ability to model each amino acid.

We also examine how successful the structural biology and structural genomics communities have been in advancing structural coverage of the *M. pneumoniae* proteome (at both "coarse" and "fine" degrees of coverage) and what role the BSGC has played. Finally, we discuss some of the remaining obstacles to obtaining complete structural coverage.

Complete data including sequences and experimental status of BSGC targets are available in the public databases TargetDB and PEPCdb [32].

METHODS

Target Selection

A structural genomics target is a protein whose structure is selected for experimental characterization. BSGC targets include *Mycoplasma* proteins as well as their homologs from other prokaryotes. In general, all rounds of target selection involved three common steps. We started each step with the set of 677 *M. pneumoniae* ORFs described in the original annotation of the genome [1]. (Note that additional ORFs have been identified more recently [33], and the current set of 687 ORFs is used throughout the remainder of this report to evaluate progress towards completion of the proteome.) Each ORF was then augmented with a family of homologs from available, fully sequenced prokaryotic genomes to make a target set. First, all target sets recognizably homologous to proteins of known structure were removed from further consideration. Next, target sets of proteins which were predicted to be unsuitable for high-throughput study (e.g., those with predicted transmembrane helices) were eliminated. Finally, specific targets were chosen from among proteins in the remaining target sets. The number of targets chosen per family, or *parallelism*, varied amongst selection rounds, as described below. A summary of methods used in different stages of target selection are shown in Table I. A typical round of target selection is described in more detail in Figure 1.

To date, there have been 6 rounds of target selection. The first round of targets, which were mainly chosen in the first year of BSGC operations, were selected using a variety of *ad hoc* methods, or because they were of interest to the BSGC experimentalists. Some aspects of this round of target selection are described elsewhere [34]. In the second round, we introduced basic standardized methods, as explained in detail below. In the third round, more sophisticated methods of detecting currently known structures were introduced, and thresholds for identifying proteins likely to be intractable for high-throughput study (e.g., length and percentage of low complexity or coiled coil) were increased in order to go beyond "low hanging fruit." In the fourth round, the parallelism was increased as BSGC experimentalists began deploying more high-throughput experimental methods, and it was noted that experimental success rates varied among similar targets from different species. In the fifth round, we chose a specialized group of targets that were more challenging to clone using automated methods. These targets presented difficulties specifically related to the genetic code used by *Mycoplasma*, as explained in detail below, but could not be ignored because more suitable homologs could not be identified. Finally, the sixth round of targets was chosen using a domain identification procedure, with the purpose of identifying tractable domain targets within full-length proteins that were set aside by filters in earlier rounds.

Identifying known structures

At the beginning of each round of target selection, all *M. pneumoniae* proteins and their homologs were considered potential targets. These were then removed from consideration if they were detectably homologous to other proteins of known structure.

Similarity to known structures was detected by first assembling a database of known protein structures, the "knownstr" database, which was updated prior to each target selection round. This database contained sequences of proteins released by PDB [35], sequences of proteins deposited in the PDB and made available while the structure is still "on hold," and sequences from TargetDB [32], for which a structure has been solved by another structural genomics center. We also included sequences of BSGC targets that have progressed to the "Traceable Map" stage, as this usually indicates the structure will soon be completed.

During each automated target selection round, sequences of all *M. pneumoniae* ORFs were compared to the knownstr database using several sequence comparison tools. PSI-BLAST [36] was used in rounds 2-6. PSI-BLAST position-specific scoring matrices (PSSMs) were constructed for each *M. pneumoniae* ORF (or predicted domain in round 6) using 10 rounds of searching our "snr" database with a matrix inclusion threshold E-value of 10^{-2} (the default value of $5x10^{-3}$ was used in round 2). The snr database included all sequences in the `swissprot`, `trembl`, and `trembl_new` files (downloaded 30 July 2001 for round 2, 30 November 2001 for round 3, 21 October 2002 for rounds 4-5, and 26 February 2004 for round 6) from Swiss-Prot [37], which had been filtered with the SEG [38] and PFILT [39] programs using default options. The filtering was done to reduce the chance of profile corruption [40], which can lead to inaccurate results. The PSSMs were used to search the knownstr database, and any hits with an E-value of 10^{-1} or below were eliminated from consideration as targets. This significance threshold was chosen to increase the likelihood of detecting more remote homologs, even though it had some risk of false positives being removed from the target list. After the second round, the matrix inclusion threshold was increased in order to increase the possibility of identifying remote homologs, at the risk of a higher rate of corrupted PSSMs.

Because of the latter possibility, we also used BLAST [41] and Pfam [42] in target selection rounds 3-6. All *M. pneumoniae* ORFs with a BLAST hit against knownstr with an E-value of 10^{-1} or below were eliminated from consideration as targets, in addition to those already eliminated by PSI-BLAST. In using Pfam to detect known structures, the HMMER tool (version 2.2g in rounds 3-5, 2.3.2 in round 6) [43] was used to compare the Pfam_ls library of hidden Markov models to both the knownstr database and the database of *M. pneumoniae* ORFs, using the family-specific "trusted cutoff" score as a cutoff for assigning significance. We eliminated from consideration all ORFs that had a significant hit to a Pfam family that had also matched at least one known structure.

Identifying targets predictably intractable for high-throughput study

As the next step in each target selection round, we eliminated *M. pneumoniae* proteins and domains that were likely to be either uninteresting or predictably intractable for high-throughput study. These included proteins with regions of amino acids predicted to be in transmembrane segments, coiled coils, regions of low complexity. We also eliminated potential targets that were long and therefore likely to be challenging; in earlier rounds (1-2) of target selection, the length cutoff was 400 amino acids, and in later rounds (3-6) it was increased to 700 amino acids. Finally, we excluded proteins annotated as ribosomal components, as these were expected to be unlikely to be stable in the absence of binding partners.

The SEG program [38] (version dated 24 May 2000) was run on all sequences to identify putative low complexity regions. Default options were used. In the round 2 of target

selection, any predicted low complexity region eliminated the ORF from consideration; in rounds 3-6, low complexity was allowed if the total length of low complexity regions did not exceed 20% of the total length of the protein (or domain in round 6).

The CCP program (written by J Kuzio at NCBI, version dated 14 June 1998), using the algorithm of Lupas [44], was used to predict coiled coil regions in all sequences. Default options were used. Thresholds for eliminating potential targets based on coiled coil predictions were the same as those used for low complexity regions (above).

Two programs were used to identify transmembrane regions. TMHMM 2.0a [45] was used, with all default options. PHDhtm [46] version 2.1 (October 1998) was also used, with the option optHtmisitMin (an option affecting the rate of false positive transmembrane predictions) set to 0.8. Any transmembrane region predicted by either protein eliminated a *M. pneumoniae* ORF from consideration as a target in rounds 2-5. In round 6, transmembrane predictions were used in assigning domain boundaries (see below).

Identifying domains

Some *Mycoplasma* ORFs that were filtered out in early selection rounds were multidomain proteins that included tractable domains of unknown structure, but had been eliminated because of homology to a single domain of known structure. Therefore, in round 6, *Mycoplasma* ORFs were divided into domains before entering the target selection filters. The procedure used was the same as that used to identify domains in the ASTEROIDS data set of the ASTRAL database [47]. Hidden Markov models of ASTRAL families and superfamilies

were used to predict domains in the *M. pneumoniae* ORFs, using the HMMER tool with a significance cutoff of 10^{-4}. BLAST was also used to compare ASTRAL sequences to all *M. pneumoniae* ORFs, using a significance cutoff E-value of 10^{-4}. Regions of *Mycoplasma* sequence matching one or more ASTRAL sequences or hidden Markov models were annotated as belonging to the same SCOP [48] superfamily as the hit with the most significant E-value produced by either method. Remaining unclassified regions were annotated using Pfam 10.0, using the Pfam_ls model library and the "trusted cutoff" score for each model to determine significance. Significant hits were annotated as Pfam domains. After Pfam annotation, remaining regions of at least 20 consecutive residues were annotated as potential unclassified domains. This procedure is identical to the one documented in the release notes for ASTRAL 1.65.

Putative domains identified by the ASTRAL procedure were further split into two parts at the end of each predicted transmembrane helix, as predicted by TMHMM 2.0a [45]. Finally, putative domains shorter than 50 residues were eliminated from further consideration as targets.

Identifying particular proteins as targets

In addition to the *M. pneumoniae* proteins themselves, homologous proteins from other prokaryotes were also chosen as targets. Each *M. pneumoniae* protein (or predicted domain in round 6) that passed through the above filters was used to search the NCBI database of proteins from sequenced bacterial and archaeal genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria), although targets were only chosen from

genomes for which the BSGC had access to purified genomic DNA. A list of all genomes from which homologous targets were chosen is given in Table S1 in the supplementary information. To find these homologs, PSI-BLAST [36] was used in rounds 2-3 (version 2.2.1) and 4-6 (version 2.2.4). PSI-BLAST PSSMs were constructed for each *M. pneumoniae* ORF using 10 rounds of searching the nonredundant sequence database "snr" (as described above) with default parameters; the PSSMs were then used to search the database of genomes. BLAST version 2.2.4 was also used (with default parameters) in rounds 4-6 to search the genome database. All proteins identified by BLAST or PSI-BLAST with E-values more significant than 10^{-4}, with the region of local similarity covering at least 50 residues, were considered as possible targets. In round 6, predicted domains from *M. pneumoniae* were used to search for possible targets. In this case, only the local region of the homologous ORF was selected as a possible target, and we also required that the region of local similarity identified by BLAST or PSI-BLAST to cover at least 80% of the length of the putative *M. pneumoniae* domain. This latter restriction was intended to decrease the possibility of selecting a fragment of a domain as a target.

Once potential targets were identified for each *M. pneumoniae* ORF or putative domain, we selected a limited number from each family as targets. The maximum number of targets chosen for each *M. pneumoniae* ORF was limited to 4 in earlier rounds (2-3) of target selection, but expanded to 10 in later rounds (4-6), after better automation became available in the BSGC experimental pipeline. Those targets were chosen as follows.

Potential targets from *M. pneumoniae* were always selected if they passed an additional screen to ensure they could be expressed in the *E. coli* expression system used at the BSGC.

M. pneumoniae [49] and other related mollicutes such as *Ureaplasma urealyticum* [50] can use UGA codons to encode the amino acid tryptophan, whereas UGA is a stop codon in *E. coli*. Thus, cloned *M. pneumoniae* proteins with this codon would express truncated proteins in *E. coli*. In cases where a UGA codon was within about 30 bases of either end of the gene, it could easily be mutated to a UGG codon during cloning, using mutating PCR primers. Other UGA codons, called internal UGA codons, could only be mutated in a more difficult multi-step cloning procedure. In rounds 1-4, targets with a maximum of 1 internal UGA codon were allowed. In round 5, this restriction was relaxed to allow 2-4 internal UGA codons. In round 6, because we wanted to clone targets using a fully automated protocol, no internal UGA codons were allowed.

The next highest priority targets to be selected were from thermophiles and halophiles, as these were expected to be experimentally more tractable, for example being partially purified by heating *E. coli* lysate [51]. These targets, if available, were chosen in order by significance of the BLAST or PSI-BLAST similarity score. If the maximum number of targets per *M. pneumoniae* ORF had not been reached after choosing these targets, additional targets were chosen from mesophilic organisms, including other paralogs from *M. pneumoniae*. These were also chosen in order by significance, with the additional restriction that the sequences had to be at least 30% identical over an aligned region of at least 50 consecutive residues. The latter restriction was intended to ensure that a reasonably accurate model could be produced for the *M. pneumoniae* ORF if the structure of the mesophile protein were to be solved. Current state-of-the-art comparative modeling methods are able to produce models of medium accuracy (about 90% of the main chain modeled to within 1.5 Å RMS error) when sequence identity between the model and the template is at least 30%;

below this threshold, alignment errors increase rapidly and become the major source of modeling error [52].

Target Deselection

Because the BSGC only seeks to solve structures for protein domains for which the structure cannot be reliably predicted via bioinformatic methods, we need to frequently check whether structures similar to our targets have been solved by other groups. We stop targets for which structures of similar proteins have been solved. Most deselection analysis steps are automated. However, the final decision on whether to stop any target is performed manually. This automated analysis and manual review are both performed weekly.

Automated analysis

The automated analysis begins with using BLAST and PSI-BLAST to compare our current target sequences to the knownstr database (described above), which is updated weekly. PSI-BLAST PSSMs are constructed for each target using 10 rounds of searching the "snr" nonredundant sequence database (described above) with a matrix inclusion threshold E-value of 10^{-2}. These PSSMs are used to search the knownstr database, and all hits with an E-value of 10^{-2} or better result in flagging the region of target sequence corresponding to the hit. BLAST hits against knownstr with E-values of 10^{-2} or better also result in flagging the region of sequence corresponding to each hit. These thresholds were chosen empirically, with the goal of being sensitive enough to detect remote homology while minimizing the time spent examining false positives.

After target residues are flagged, those proteins that possess at least one region of 50 consecutive residues not flagged by hits are automatically left in the structural genomics pipeline. This is because even if some parts of a target are found to be similar to proteins of known structure, the remaining region may potentially contain a domain for which no reliable prediction of structure could be made through the bioinformatic methods used. Targets that are similar to proteins of known structure over virtually their entire length (without a stretch of 50 consecutive residues not flagged by hits) are identified for manual review to determine whether these targets should be deselected.

Manual review of target deselection candidates

Because the bioinformatic procedure above may result in false positives, targets identified by the procedure are manually examined to determine if work should be stopped. The decision about whether to stop work on a target is made by the experimentalists working on the target and reflects a cost-benefit analysis of how much work would be required to finish the structure versus the potential for new information to be gained. This decision is informed by the degree of sequence similarity with the known structure(s) and implications for accuracy of a comparative homology model, and whether functionally important residues in the known structure are conserved in the target. Generally, targets that have been crystallized are not deselected when the structure of a similar protein has been solved, because after data collection, the target structure may easily be solved using molecular replacement. If a target has not been purified, it is generally stopped if the fold prediction is thought to be reliable, even if the similarity is insufficient to allow accurate

modeling. Targets that have been purified but not crystallized are usually stopped only if an accurate model can be constructed, and if crystallization trials are proceeding poorly.

After a decision is made on whether to deselect or continue work on the target, the decision is recorded. If the target is continued, it is not recommended again for deselection by the automatic procedure unless a new structure is solved that is identified as similar to the target. We are identifying ways to automatically perform much of the review process, in order to more quickly process larger numbers of targets.

Quantifying coverage of Mycoplasma *proteomes*

The ultimate goal of structural genomics is to provide structural information for the complete repertoire of biological macromolecules. In this report, we measure progress towards that goal as "coverage," the fraction of sequences or residues in a set (such as a proteome) for which structural information is available or can be inferred. If a region of sequence is at least 30% identical to a protein with experimentally determined three-dimensional structure, the region is considered covered at a "fine" level. If homology is detectable, regardless of sequence identity, the region is considered covered at a "coarse" level. Details of these calculations are described below.

Per-sequence coverage of a proteome was measured as the fraction of sequences in the proteome that have at least one region covered by structural annotation. Per-residue coverage was calculated by dividing the number of residues covered by structural annotations by the total number of residues. In the latter case, all residues between the

endpoints of a local alignment (e.g., from BLAST or Pfam) were treated as covered by the annotation, whether they are aligned to a residue or a gap. We also estimated the per-residue coverage of regions of the proteome predicted to be "HT-tractable and interesting" when using high-throughput (HT) experimental methods for structure determination. For this calculation, we excluded regions predicted to be transmembrane, low complexity, or coiled coil, as well as short interstitial regions (fewer than 50 residues) between predicted transmembrane regions and regions of structural annotation. The actual number of such residues in each proteome varies slightly in each calculation, as the interstitial regions change depending on which regions are annotated as matching a domain of known structure. For example, there are more regions annotated as covered at a coarse level than at a fine level, so there are additional residues in short interstitial regions in the latter calculation. However, in general, the number of predicted HT-tractable and interesting residues is about 85% of the total number of residues in each proteome. Predictions of low complexity, coiled coil, or transmembrane regions were performed during target selection, as described above. We report both variants of per-residue coverage in tables.

Our analysis of coverage is based on an updated annotation of the *M. pneumoniae* genome [33], which includes 687 proteins and 239,722 residues. We also measured coverage of the *M. genitalium* proteome [2], which is annotated as containing 486 proteins and 175,930 residues.

As a baseline, we calculated coverage of the *M. pneumoniae* and *M. genitalium* proteomes by known structures prior to the establishment of the BSGC on 1 September 2000. We then measured coverage by structures solved by the BSGC, as well as coverage that

would result if structures of targets selected in each round of target selection were successfully completed. Finally, we measured coverage by all current structures (as of 13 July 2004) in order to determine the relative impact of the BSGC's efforts.

Coarse coverage was evaluated using BLAST (2.2.4), PSI-BLAST (2.2.4), and Pfam 10.0. BLAST was used with default parameters to search each *M. pneumoniae* ORF against the knownstr database and a database of BSGC targets. A PSI-BLAST PSSM was constructed for each *M. pneumoniae* and *M. genitalium* ORF using 10 rounds of searching the snr nonredundant sequence database (as described above, downloaded 26 February 2004) with default parameters; the PSSMs were then used to search the knownstr database and the database of BSGC targets. An E-value cutoff of 10^{-4} was used as a threshold for evaluating significance for both BLAST and PSI-BLAST; for PSI-BLAST, this corresponds to about a 1% error rate in genome annotation [53,54]. The HMMER tool (version 2.3.2) [43] was used to compare the Pfam_ls library of hidden Markov models from Pfam 10.0 to the knownstr database, the database of *M. pneumoniae* and *M. genitalium* ORFs, and the database of BSGC targets, using the family-specific "trusted cutoff" score as a cutoff for assigning significance. Local regions of these sequences were assigned as matching each other if they both had significant matches to the same Pfam family.

Fine coverage was evaluated using the subset of coarse coverage results produced by BLAST and PSI-BLAST for which the percentage identity calculated by (PSI-) BLAST was above 30% in the region of alignment.

RESULTS

Experimentally difficult regions of Mycoplasma *proteomes*

The focus of BSGC effort is on aspects of the *M. pneumoniae* proteome which are both interesting and tractable to high-throughput (HT) methods of structure determination. This encompasses the whole proteome of 687 ORFs, excluding all regions predicted to span the membrane, coiled coil regions, short loops between domains, and low complexity regions. Of the 687 ORFs in *M. pneumoniae*, 149 (21.7%) have at least one predicted transmembrane helix. 33 of 687 proteins (4.8%) have at least 20% of their sequence predicted as coiled coil, and 43 of 687 proteins (6.3%) have at least 20% of their sequence predicted as low complexity. A total of 201 of 687 proteins (29.3 %) were considered intractable to high-throughput study due to meeting at least one of these three criteria. A total of 14.8% of the residues in the proteome (35,419/239,722) are in regions predicted to be either low complexity, coiled coil, or transmembrane helix, and thus either uninteresting or experimentally difficult to solve using high-throughput methods of structure determination. An additional 3,133 residues (1.3%) in the proteome are in short (<50 residue) interstitial regions between transmembrane helices and currently known structures (at the coarse level of similarity). The percentages are similar for *M. genitalium*. Of 486 ORFs, 111 (22.8%) have at least one predicted transmembrane helix, 19 (3.9%) have at least 20% predicted coiled coil, and 22 (4.5%) have at least 20% predicted low complexity. A total of 136 of 486 proteins (28.0%) were considered intractable to high-throughput study due to meeting at least one of the three criteria. A total of 14.1% of *M. genitalium* residues (24,880/175,930) are

in regions predicted to be low complexity, coiled coil, or transmembrane helix, and an additional 738 residues (0.5%) are in the short interstitial regions described above.

Coverage by BSGC targets

Coverage of the *M. pneumoniae* and *M. genitalium* proteomes by structures released prior to the establishment of the BSGC on 1 September 2000, and by targets in each round of target selection to date, are shown in Table II. Only 142 of 687 *M. pneumoniae* proteins (20.7%) and 20.5% (39,448/192,673) of the predicted HT-tractable and interesting residues could be accurately modeled based on structures available prior to BSGC establishment. More than twice as many--297 of 687 proteins (43.2%), or 43.1% (84,324/195,732) of the HT-tractable and interesting residues--could be reliably assigned to a fold at that time. A higher fraction of *M. genitalium* proteins were covered: 137 of 486 proteins (28.8%) and 26.6% (37,855/142,422) of HT-tractable and interesting residues could be modeled, while 262 of 486 proteins (53.9%) and 52.4% (75,936/144,943) of HT-tractable and interesting residues could be reliably assigned to a fold.

The first round of preliminary and manually selected targets produced the greatest incremental increases in coverage. However, the parallelism in this target set was low: an average of only one to two targets were selected for each *Mycoplasma* protein of interest.

The next three sets of automatically selected targets each provided incremental improvements in coverage, as well as a deliberate increase in the parallelism in the pipeline. In rounds 2-3, up to four targets were chosen for each *M. pneumoniae* protein of interest,

counting targets already chosen in other rounds and to cover other *M. pneumoniae* proteins. This increased the average number of targets per protein to more than 3, although there were some cases where fewer than four homologs could be found that met our criteria to be targets. In cases where multiple paralogs of a gene existed within *M. pneumoniae*, the number of targets per *Mycoplasma* ORF was sometimes more than 4, as targets chosen to cover one paralog might also be similar to others. In round 4, the maximum number of targets chosen per *M. pneumoniae* protein was increased to 10. However, this did not increase the actual redundancy in the pipeline as much as expected, as nearly all available homologs meeting our criteria as targets had already been chosen.

In the 4th round of target selection, 65 potential targets were eliminated by the filter that prevented targets with more than 1 internal UGA codon from being chosen. However, 46 of these rejected targets were *M. pneumoniae* proteins with no more tractable homologs in our dataset. In round 5, the UGA codon limit was relaxed from 1 to 4 internal UGA codons permitted in order to target some of these proteins using a more complex multi-step cloning approach to mutate each of the codons to UGG. While 33 of the 46 previously rejected targets were selected in this round, the other 13 had between 5 and 21 internal UGA codons, so were judged to be too difficult for this technique to succeed in a manner suitable for structural genomics. The 33 targets chosen led to a significant increase in coverage of *M. pneumoniae*: 37-43 more proteins and 10-12% more residues depending on whether coverage is measured at the coarse or fine level. This step had a smaller impact on coverage in *M. genitalium* (only 13 more proteins) as most targets chosen in round 5 were unique to *M. pneumoniae*.

In the 6th round of target selection, individual predicted domains were selected instead of full length targets, in order to increase the number of potential tractable targets. Domain prediction resulted in greater coverage of the *Mycoplasma* proteomes as well as more than doubling the parallelism in the experimental pipeline. We expect some failures of these targets due to inaccurate prediction of domain boundaries: a preliminary analysis based on successive versions of SCOP showed that the domain prediction method accurately predicts 65% of the domain boundaries to within 10 residues of the manually assigned boundaries in SCOP, and 80% of the boundaries are correctly predicted within 20 residues (unpublished). In addition, some domains are unable to fold on their own, even if the boundaries are correctly identified. However, the increased parallelism in the pipeline should partially alleviate these potential problems. Preliminary experimental success rates for these targets are reported as supplementary information.

Mycoplasma *residues remaining uncovered by targets*

After 6 rounds of target selection, current BSGC targets cover 550 of 687 *M. pneumoniae* proteins (80.1%) and 78.7% (161,281/204,812) of the HT-tractable and interesting residues at the coarse level. The remaining regions not covered by BSGC targets form 230 continuous stretches of sequence at least 50 residues long. Of these, 121 contain 1 or more internal UGA codons, so were not chosen as targets during the last round of target selection. These may be selected in future rounds of target selection, as the UGA problem may be solved by using other expression systems or by cloning homologs from other bacteria. The other 109 regions contain more than 20% predicted coiled coil or low complexity regions, or at least one transmembrane helix, which would prevent them from being chosen as targets

under our current criteria. While the coiled coil or low complexity residues in each region are not considered "HT-tractable and interesting," the other residues in each region are. One of these regions is the ribosomal protein S21, which was excluded due to potential inability to fold in the absence of binding partners, but which is not part of current ribosomal structures. The remaining 109 regions may prove to be intractable to high throughput studies.

Of the 687 *M. pneumoniae* proteins, 223 (32.5%) have no homologs outside of other *Mycoplasma* and *Ureaplasma* species, and 54 (7.9%) are ORFans [23], having no homologs outside *M. pneumoniae*. Of the 230 remaining regions in *M. pneumoniae* not covered by targets, 83 (36%) are in proteins that have no homologs outside of other *Mycoplasma* and *Ureaplasma* species, and 14 (6.1%) are in ORFans. Therefore, the remaining regions do not appear to be biased towards ORFans. For most of the 121 regions currently not selected due to UGA codons, it is likely that targets may be chosen from other species when additional genomic DNA becomes available.

Current structural coverage, and impact of BSGC

As shown in Table III, coarse structural coverage of the *M. pneumoniae* proteome has increased from 297 of 687 proteins (43.2%) in 1 September 2000 to 391 of 687 proteins (56.9%) due to the solution of experimental structures since the start of the BSGC. Coverage measured as a fraction of interesting and HT-tractable residues has increased over the same time period from 43.1% (84,324/195,732) to 59.4% (119,433/201,170). Fine coverage has increased from 142 of 687 proteins (20.7%) to 254 of 687 proteins (37.0%), or from 20.5%

(39,448/192,673) to 36.7% (71,405/194,362) of the interesting and HT-tractable residues. This represents a near doubling of fine coverage, as well as a significant increase in coarse coverage.

To date (as of 13 July 2004), the BSGC has solved 69 structures of 51 different targets (some of the structures are for the same targets, under different conditions or with bound ligands). A disproportionate number of the solved structures to date have been from thermophiles (32 of 51 solved targets, or 63%, versus 284 of 945 total targets, or 30%), which were usually selected to cover *M. pneumoniae* proteins at a coarse rather than fine level. Therefore, BSGC structures have had more of an impact on coarse coverage of the proteome than on fine coverage. The relative impact of BSGC structures on coverage of *Mycoplasma* proteomes is illustrated in Figure 2. Coarse coverage of *M. pneumoniae* has increased by 29 proteins (4.3% of the 687 proteins in the proteome) due to BSGC structures, while increasing by 83 proteins (12.1%) due to all non-BSGC structural genomics and structural biology efforts over the same time period. There is significant overlap between the two groups: targets similar to 18 *M. pneumoniae* proteins were solved by both BSGC and non-BSGC groups. In 13 of these 18 cases, the BSGC solved and released the target structure prior to the other groups. However, even under the assumption that structures similar to all 18 *M. pneumoniae* proteins would have been solved in the absence of the BSGC, the 11 *M. pneumoniae* proteins covered by targets solved only at the BSGC account for 11.7% (11 of 94 proteins) of the total increase in coarse coverage. The 24 structures solved either solely or first at the BSGC account for 25.5% (24 of 94) of the total increase in the number of proteins with coarse coverage over the lifetime of the BSGC to date. Similarly, coarse coverage of *M. genitalium* has increased from 262 proteins (53.9% of the 486 proteins in the proteome) to 348 proteins (71.6%).

Coverage of interesting and HT-tractable residues in *M. genitalium* increased from 52.4% (75,936/144,943) to 72.0% (108,155/150,312). BSGC efforts account for coverage of 25 *M. genitalium* proteins, 16 of which were also covered by structures solved elsewhere (although 12 of the 16 were first covered by BSGC structures). The 9 proteins for which targets were solved only at the BSGC represent 10.5% (9 of 86) of the total increase in coarse coverage of *M. genitalium* over the lifetime of the BSGC, while the 21 proteins solved either solely or first at the BSGC account for 24.4% (21 of 86) of the total increase in *M. genitalium* proteins covered.

While fine coverage of both *Mycoplasma* proteomes increased by a smaller amount due to BSGC structures, there was less overlap with structures solved by other groups. Fine coverage of *M. pneumoniae* has increased by 20 proteins (2.9% of the 687 proteins in the proteome) due to BSGC structures, while increasing by 98 proteins (13.2%) due to all other structures solved over the same time period. Only 6 proteins overlap between the two groups, and in 4 of these 6 cases, the BSGC solved the target prior to the other groups. The 14 *M. pneumoniae* proteins covered only by BSGC structures account for 12.5% (14 of 112) of the total increase in fine coverage of the proteome, and the 18 proteins covered solely or first by BSGC targets account for 16.1% (18 of 112) of the increase. Fine coverage of *M. genitalium* has increased from 137 proteins (28.2% of the 486 proteins in the proteome) to 243 proteins (50%) over the lifetime of the BSGC. Coverage of the interesting and HT-tractable residues in *M. genitalium* has increased from 26.6% (37,855/142,422) to 47.1% (67,9705/144,024) during the same time period. BSGC efforts account for coverage of 17 *M. genitalium* proteins, 7 of which were also covered by structures solved elsewhere (5 of the 7 were first covered by BSGC structures). The remaining 10 proteins represent 9% (10 of 106)

of the total increase in fine coverage of *M. genitalium* over the lifetime of the BSGC; proteins solved solely or first by the BSGC account for 14.2% (15 of 106) of the increase.

It is interesting to contrast the increased coverage of *Mycoplasma* provided by BSGC structures with coverage provided by one of the most impressive structural biology achievements made at about the same time the BSGC was getting underway: high-resolution structures of the ribosome [55-57]. Some individual ribosomal proteins had been solved prior to the first of these studies, and these prior structures contributed to fine coverage of 17 *M. pneumoniae* proteins (2.5% of the 687 proteins in the proteome) and coarse coverage of 21 proteins (3.1% of the proteome). Ribosomal structures currently contribute to fine coverage 47 proteins in *M. pneumoniae* (6.8% of the 687 proteins), and coarse coverage of 57 proteins (8.3%). Currently, all annotated ribosomal proteins in *M. pneumoniae* except L33 type 2, L28, and S21 are covered at least coarsely. While ribosomal structures have had a greater impact on coverage than all structures solved at the BSGC, it is unlikely that any single macromolecular complex that is studied in the future will provide such an increase.

Impact of target deselection

As of 1 June 2004, 324 separate target deselection recommendations had been issued by the automated system, an average of 2.4 per week since the system was deployed in October 2001. 146 of the suggestions were overridden, and 178 were followed, resulting in stopping work on a target. Recommendations are automatically cancelled and re-issued if additional structural information becomes available prior to the recommendation being acted on, and these statistics do not include hundreds of such cases: multiple recommendations

before action was taken were treated as a single recommendation. Many of the recommendations that were issued concerned the same targets: the 146 overridden suggestions were issued on a total of 54 targets, and 30 of these targets were eventually stopped after two or more deselection recommendations. Overall, recommendations were issued on 202 separate targets, of which 178 were deselected.

Most of the target deselection recommendations took place prior to the last round (round 6) of target selection on 22 March 2004, at a time when there were fewer than 400 targets being actively worked on (neither stopped nor solved). As there are currently almost 700 active targets, we expect the number of recommendations to increase accordingly. Figure 3 shows the percentage of targets that were deselected over time, as a fraction of the cumulative number of targets chosen. Figure 4 shows the stages at which targets were stopped: 49 of 178 (27.5%) were stopped after the target protein was purified. About half (86 of 178, or 48.3%) of the targets were stopped because we solved a "parallel" target, and about the same number (87 of 178, or 48.9%) were stopped due to another structural genomics center or structural biology laboratory solving a structure. Only five targets have been stopped solely due to experimental difficulty, although experimental difficulty is a factor taken into consideration during the manual review phase of target deselection.

65 of the 178 deselected targets (37%) were stopped based on the sequence of a homologous protein being released by the PDB, at the time of either the deposition or release of the structure. In 13 of these cases, the recommendation to stop was based on a structure that was on hold and unavailable to us, but for which the sequence was available prior to the release of the structure. In these cases, the time between release of the sequence and release

of the structure by the PDB ranged from 33 to 396 days, with an average hold time of 231.9 days. In these cases, the crystallographers' decision to release their sequences at the time of deposition allowed us to stop these targets almost 8 months earlier on average than we could have if the sequences had not been made available. In the other 52 cases, the sequence was not made available until the structure was also released. In these cases, the hold times (time between deposition and release of the structure and sequence) ranged from 19 to 1,515 days, with an average hold time of 151.4 days. Had the sequences of these 52 structures been made available at the time deposition to the PDB, the deselection recommendations could have been made almost 5 months earlier on average (and in the longest case, 1QGD, in which the structure was on hold for over four years, the BSGC targets would not have been selected).

To evaluate the impact of stopping work on 178 targets, we measured incremental coverage of the *M. pneumoniae* proteome at coarse and fine levels that would have resulted had the targets been solved, relative to the actual current coverage. At a fine level, coverage would have been increased by 19 proteins (2.7% of the 687 proteins in the proteome, or 2.5% of interesting and HT-tractable residues), and at a coarse level, coverage would have been increased by only 1 protein (0.1% of proteins, or 0.2% of the interesting and HT-tractable residues). This is not surprising, as the target deselection procedure focuses on remote homology; if finishing a target would lead to more coverage at a fine level but not at a coarse level, the target is usually stopped.

Impact on coverage of other proteomes

One of the secondary goals of choosing a minimal proteome as the focus of structural genomics efforts at the BSGC was to evaluate the impact on coverage of larger proteomes. The idea is that a minimal proteome is a ubiquitous proteome, and that the complete structural complement of a minimal proteome would serve as a platform for understanding larger proteomes [15]. In an earlier Pfam-based study [30], we showed that maximum coverage across multiple species is obtained by solving structures from large families; solving structures of proteins not classified in large Pfam-A families has little impact on coverage of other species. We used HMMER [43] to identify all Pfam-A (version 10.0) families in our solved targets, using the "trusted cutoff" for each family as a measure of determining significance. Three of our solved targets had no hits in Pfam-A, and may represent small families restricted to a few bacteria. Pfam-A families for which the BSGC solved the first structure are shown in Table IV. All but two of these 24 families are larger than the median family size (36) in Pfam 10.0.

Using methods described elsewhere [30], we measured coverage in several other proteomes, as well as Swiss-Prot and TrEMBL. Results are shown in Table V. Most of the 24 Pfam families match at least one family in each proteome; the total number of hits ranges from 20 in the *M. jannaschii* proteome to 100 in *A. thaliana*. Overall, the 24 families hit a total of 1,122 proteins in Swiss-Prot (from Pfamseq 10.0) and 3,737 in Swiss-Prot and TrEMBL combined. Thus, the families solved first or only at the BSGC are in fact nearly ubiquitous across a variety of commonly studied eukaryotic and prokaryotic proteomes. Note that

BSGC structures added approximately 1% to the number of proteins covered in other prokaryotes such as *E. coli*, *M. jannaschii*, and *M. tuberculosis*.

Cellular functions of targets

One of the goals of structural genomics is to study proteins of unknown function and "hypothetical proteins," as the three-dimensional structures of these proteins often suggest biochemical or biophysical functions [58,59]. Biochemical and cellular functions of microbial proteins are annotated in the Comprehensive Microbial Resource [60]. The annotated functions of all *M. pneumoniae* and *M. genitalium* proteins, and our targets, are shown in Table VI.

As shown in Table VI, the majority of our targets (508 of 945, or 54%) are annotated as hypothetical proteins, unclassified function, unknown function, or not annotated. Proteins in these categories also constitute the majority of *M. pneumoniae* proteins (363 of 687, or 53%) and a large fraction of *M. genitalium* proteins (195 of 486, or 40%). Proteins in this set have relatively little structural coverage: only 35% of these *M. pneumoniae* proteins (127 of 363) and 50% of these *M. genitalium* proteins (97 of 195) are covered by current structures at a coarse level. Only cellular envelope proteins (50 in *M. pneumoniae* and 28 in *M. genitalium*) have less coverage, as expected since many of these proteins contain transmembrane regions. Although TIGR role annotations were not explicitly considered when choosing targets, this analysis shows that most currently active targets correspond to roles which have the least amount of current structural coverage.

Current active targets

As of 13 July 2004, there are 649 current active targets (targets that have not been solved or stopped), as shown in Table I. The distribution of experimental stages of these targets is shown in Table S2 and discussed further in the supplementary material. Of these, the vast majority (459 of 649, or 71%) were selected in the most recent round, round 6, several months before. In the prior three automated sets (rounds 2-4), approximately half of the targets (115/227) are still active, the remaining targets having been solved or stopped due to homology with a solved structure. The overall fraction of targets for which the BSGC has solved a structure in these three rounds is approximately 8% (11% in round 2, 5% in round 3, and 8% in round 4, or 19/227 overall). The fraction of solved targets is slightly higher in round 2, as expected since these targets have been active for the longest time. No targets in the final two rounds (5-6) have been solved, as they have only been active for a few months. The first round has a much higher fraction of solved targets: structures for 32 of 163 targets (20%) were solved. We suspect this is due to two factors. First, these targets have been in the experimental pipeline for longer, so there has been more time to work around experimental difficulties in a "multi-path" approach [61]. Second, these targets include some targets manually selected by experimentalists as interesting, and a share of the work in these cases was done by collaborators, allowing more attention to be focused on these targets. The expected rate at which full-length targets will be solved in the future therefore probably lies somewhere between the 11% observed for round 2 and the 20% observed for round 1. Because many targets in these two rounds were deselected due to a homolog being solved at the BSGC or elsewhere (83 of 163, or 51%, in round 1, and 38 of 92, or 41%, in round 2), this fraction of targets which have been solved represents a lower bound on the

percentage of targets which are tractable using our current methods. We expect the fraction of solved structures for predicted domain targets to be somewhat lower than for full length targets, both because the targets themselves are expected to be relatively more difficult experimentally (for reasons described above) and because the parallelism in round 6 is higher, so more will be deselected as a result of solving a parallel target.

DISCUSSION

We have documented the methods of target selection and deselection deployed to date at the BSGC, demonstrating an evolving strategy that started with "low-hanging fruit:" targets that are most likely to be tractable, and least similar to currently known structures. In successive rounds of target selection, both more experimentally challenging targets as well as targets more similar to known structures were selected for experimentation. We also succeeded in increasing the parallelism of targets in our pipeline, in response to reports that homologous proteins may exhibit very different degrees of tractability. In practice, this appears to have been effective: targets that were deselected because we solved a parallel target were at a variety of stages at the time one of the parallel targets was solved.

Our target deselection procedure has been very efficient in preventing the BSGC from spending effort on targets that would result in little incremental coverage of *Mycoplasma* proteomes. However, a drawback of the procedure is that it requires a significant amount of human effort to manually examine new recommendations every week. As we expect the required effort will scale almost linearly with the number of active targets, structural

genomics centers such as the BSGC will need to further automate target deselection as the overall throughput of structural genomics increases.

The automated procedure for recommending target deselection relies on timely availability of the sequences of newly solved structures. One of the primary sources of data is the sequences of "on hold" structures from the PDB. Upon deposition of a new structure, the authors of a PDB entry may choose whether to make the sequence available immediately or hide the sequence until release of the structure. Of the 2722 structures awaiting release today (17 August 2004), the sequence is available for only 935; for those structures held for publication (1691 structures) or release on a future date (332), sequences are available for less than half (883/2022, or 44%). More timely access to the remaining sequences, or the ability to compare structural genomics target sequences to hidden "on-hold" sequences, would enable more efficient use of resources by the BSGC and other structural genomics centers.

Our primary goal in target selection was coverage of the tractable and interesting portions of the *M. pneumoniae* proteome at a coarse level of similarity. If all our current targets were solved, either at the BSGC or by the structural biology community, we would be approximately 80% of the way towards achieving that goal. Of the remaining 20%, we estimate that approximately half could be targeted with high-throughput methods, if the procedure for introducing multiple point mutations during cloning were to be fully automated. The remaining 10% of the proteome that has not been targeted to date consists of tractable and interesting regions closely linked to experimentally problematic regions such as low complexity or transmembrane regions, and therefore may prove more resistant to high-throughput methods. It is also unlikely that all current targets in the pipeline are

actually tractable to high-throughput study, as some targets may be unstructured in the absence of a required partner or ligand.

Our focus on coarse coverage of the proteome has led to an impressive increase in coverage with a relatively modest number of solved structures. In the nearly 4 years since September 2000, over 8,000 structures have been deposited to the PDB. While the 69 structures contributed by the BSGC account for less than 1% of that total, these structures account for approximately 25% of the total incremental increase in coarse structural coverage of the *M. pneumoniae* and *M. genitalium* proteomes during that time. Structures solved by the BSGC include the only structural representatives for 10 Pfam-A families, and were the first structural representatives for 14 additional Pfam-A families. These families are nearly ubiquitous across a wide variety of eukaryotic and prokaryotic proteomes.

ACKNOWLEDGEMENTS

We are grateful to all structural biologists who promptly reported sequences and structures to the PDB. We thank all the BSGC experimentalists for their work. Michael Levitt inspired some early ideas in this work. In-Geol Choi and Igor Grigoriev performed bioinformatic analyses for the first round of target selection. Some computers used in target selection calculations were purchased through an IBM SUR grant. This work is supported by grants from the NIH (1-P50-GM62412 and 1-K22-HG00056), the Searle Scholars Program (01-L-116), and the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

REFERENCES

1. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 1996;24(22):4420-4449.
2. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al. The minimal gene complement of Mycoplasma genitalium. Science 1995;270(5235):397-403.
3. Koonin EV. How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet 2000;1:99-116.
4. Peterson SN, Hu PC, Bott KF, Hutchison CA, 3rd. A survey of the Mycoplasma genitalium genome by using random sequencing. J Bacteriol 1993;175(24):7918-7930.
5. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC. Global transposon mutagenesis and a minimal Mycoplasma genome. Science 1999;286(5447):2165-2169.
6. Koonin EV, Mushegian AR, Rudd KE. Sequencing and analysis of bacterial genomes. Curr Biol 1996;6(4):404-416.
7. Ouzounis C, Casari G, Valencia A, Sander C. Novelties from the complete genome of Mycoplasma genitalium. Mol Microbiol 1996;20(4):898-900.
8. Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol 1998;1(1):55-67.
9. Brenner SE. Errors in genome annotation. Trends Genet 1999;15(4):132-133.
10. Balasubramanian S, Schneider T, Gerstein M, Regan L. Proteomics of Mycoplasma genitalium: identification and characterization of unannotated and atypical proteins in a small model genome. Nucleic Acids Res 2000;28(16):3075-3082.
11. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287(4):797-815.
12. Rychlewski L, Zhang B, Godzik A. Fold and function predictions for Mycoplasma genitalium proteins. Fold Des 1998;3(4):229-238.
13. Chandonia JM, Cohen FE. New local potential useful for genome annotation and 3D modeling. J Mol Biol 2003;332(4):835-850.
14. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2(10):801-809.
15. Kim SH. Structural genomics of microbes: an objective. Curr Opin Struct Biol 2000;10(3):380-383.
16. Matte A, Sivaraman J, Ekiel I, Gehring K, Jia Z, Cygler M. Contribution of structural genomics to understanding the biology of Escherichia coli. J Bacteriol 2003;185(14):3994-4002.
17. Goulding CW, Apostol M, Anderson DH, Gill HS, Smith CV, Kuo MR, Yang JK, Waldo GS, Suh SW, Chauhan R, Kale A, Bachhawat N, Mande SC, Johnston JM, Lott JS, Baker EN, Arcus VL, Leys D, McLean KJ, Munro AW, Berendzen J, Sharma V, Park MS, Eisenberg D, Sacchettini J, Alber T, Rupp B, Jacobs W, Jr., Terwilliger TC. The TB structural genomics consortium: providing a structural foundation for drug discovery. Curr Drug Targets Infect Disord 2002;2(2):121-141.
18. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L,

Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. Protein Sci 2002;11(4):723-738.
19. Liu J, Rost B. Target space for structural genomics revisited. Bioinformatics 2002;18(7):922-933.
20. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8(6):559-566.
21. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B. Automatic target selection for structural genomics on eukaryotes. Proteins 2004;56(2):188-200.
22. Burley SK, Bonanno JB. Structural genomics. Methods Biochem Anal 2003;44:591-612.
23. Fischer D. Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. Protein Eng 1999;12(12):1029-1030.
24. Gaasterland T. Strategies for structural genomics target selection. ScientificWorldJournal 2002;2 Suppl 2:67.
25. Watson JD, Todd AE, Bray J, Laskowski RA, Edwards A, Joachimiak A, Orengo CA, Thornton JM. Target selection and determination of function in structural genomics. IUBMB Life 2003;55(4-5):249-255.
26. Mittl PR, Grutter MG. Structural genomics: opportunities and challenges. Curr Opin Chem Biol 2001;5(4):402-408.
27. Brenner SE. Target selection for structural genomics. Nat Struct Biol 2000;7 Suppl:967-969.
28. Blundell TL, Mizuguchi K. Structural genomics: an overview. Prog Biophys Mol Biol 2000;73(5):289-295.
29. Brenner SE, Levitt M. Expectations from structural genomics. Protein Sci 2000;9(1):197-200.
30. Chandonia JM, Brenner SE. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins 2005;58(1):166-179.
31. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R. Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. Nucleic Acids Res 1997;25(4):701-712.
32. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. Bioinformatics 2004.
33. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, Doerks T, Sanchez-Pulido L, Snel B, Suyama M, Yuan YP, Herrmann R, Bork P. Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Res 2000;28(17):3278-3288.
34. Grigoriev IV, Choi IG. Target selection for structural genomics: a single genome approach. Omics 2002;6(4):349-362.
35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-3402.
37. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31(1):365-370.

38. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18(3):269-285.
39. Jones DT, Swindells MB. Getting the most from PSI-BLAST. Trends Biochem Sci 2002;27(3):161-164.
40. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29(14):2994-3005.
41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-410.
42. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32 Database issue:D138-141.
43. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-763.
44. Lupas A. Prediction and analysis of coiled-coil structures. Methods Enzymol 1996;266:513-525.
45. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305(3):567-580.
46. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995;4(3):521-533.
47. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32 Database issue:D189-192.
48. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536-540.
49. Inamine JM, Ho KC, Loechel S, Hu PC. Evidence that UGA is read as a tryptophan codon rather than as a stop codon by Mycoplasma pneumoniae, Mycoplasma genitalium, and Mycoplasma gallisepticum. J Bacteriol 1990;172(1):504-506.
50. Blanchard A. Ureaplasma urealyticum urease genes; use of a UGA tryptophan codon. Mol Microbiol 1990;4(4):669-676.
51. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001;65(1):1-43.
52. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294(5540):93-96.
53. Muller A, MacCallum RM, Sternberg MJ. Benchmarking PSI-BLAST in genome annotation. J Mol Biol 1999;293(5):1257-1271.
54. Brenner SE. Molecular propinquity: evolutionary and structural relationships of proteins. Cambridge: Cambridge University; 1996.
55. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 2000;289(5481):905-920.
56. Wimberly BT, Brodersen DE, Clemons WM, Jr., Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. Structure of the 30S ribosomal subunit. Nature 2000;407(6802):327-339.
57. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. Cell 2000;102(5):615-623.

58. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci U S A 1998;95(26):15189-15193.
59. Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S, Kim R. Structure-based functional inference in structural genomics. J Struct Funct Genomics 2003;4(2-3):129-135.
60. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The Comprehensive Microbial Resource. Nucleic Acids Res 2001;29(1):123-125.
61. Kim SH, Shin DH, Liu J, Oganesyan V, Chen S, Xu Q, Kim JS, Das D, Schulze-Gahmen U, Holbrook S, Holbrook ER, Martinez B, Oganesyan N, DeGiovanni A, Lou Y, Hernrizuez M, Huang CC, Jancarik J, Pufan R, Choi IG, Chandonia JM, Hou J, Gold B, Yokota H, Brenner SE, Adams PD, Kim R. Structural Genomics of Minimal Organisms and Protein Fold Space. J Struct Funct Genomics 2005;in press.

FIGURE LEGENDS

Figure 1: Details of round 4 of target selection. The number of *M. pneumoniae* ORFs eliminated by each filter is shown, and also expressed as a percentage of the number of targets entering the filter. The final filter, for UGA codons, eliminated only the *M. pneumoniae* ORF but not other members of the family.

Figure 2: Percentage of *Mycoplasma* proteins covered at the coarse level by pre-BSGC, BSGC, and non-BSGC targets. A timeline illustrates the relevant dates of PDB deposition. Detailed data, including fine coverage and per-residue coverage, is given in Table III. Eight structures solved prior to the formal establishment of the BSGC which were selected as BSGC targets in round 1 are included as BSGC targets rather than pre-BSGC targets, even though they were deposited into the PDB prior to 1 September 2000.

Figure 3: Percentage and number of BSGC targets that have been stopped, over time. The percentage stopped is calculated as a fraction of the total number of targets that had been selected prior to each date.

Figure 4: Target stage at time of deselection, for the 178 deselected targets. Five targets were deselected due to experimental difficulty, 86 because the BSGC solved a homologous target, and 87 because the structure of a homologous protein was solved elsewhere.

Table I: Methods used in BSGC target selection rounds. The number of targets selected in each round is given in parentheses next to the description of the round, followed by the numbers solved and active (neither solved nor stopped) as of 13 July 2004. The "Max Targets per MP" column refers to the maximum number of protein targets selected for each *M. pneumoniae* protein that met the criteria for that round. In round 5, the maximum number of targets per *M. pneumoniae* ORF was theoretically limited to 10 as in round 4, but was actually 1 since these proteins did not have homologs in other bacteria.

Round (Date Selected)	**Description (number of targets / number solved / currently active)**	**Method of detecting known structure**	**Standard for eliminating less tractable proteins**	**Max Targets per MP**
1 (various dates)	Preliminary and manually selected targets (163/32/42)	*ad hoc*	*ad hoc*	*ad hoc*
2 (28 Aug 2001)	First automated set (92/10/44)	PSI-BLAST (v. 2.2.1, snr dated 30 July 2001, h=0.005, $e=10^{-4}$)	Any predicted coiled coil, low complexity, and transmembrane regions. Length > 400 AA. For *Mycoplasma* genes, max of 1 internal UGA codon.	4
3 (25 Feb 2002)	Second automated set (42/2/28)	Pfam (v. 7.0, trusted cutoff), BLAST (v. 2.2.1, $e=10^{-1}$), PSI-BLAST (v. 2.2.1, $h=10^{-2}$, snr dated 30 Nov 2001, $e=10^{-1}$)	Same as #2, but max length increased to 700, and thresholds for predicted coiled coil and low complexity regions raised to 20%.	4
4 (7 Nov 2002)	Third automated set (93/7/43)	Same as #3	Same as #3	10
5 (3 Mar 2004)	Multi-UGA targets (33/0/33)	Same as #3	Same as #3, but allow 2-4 internal UGA codons	10 (1)
6 (22 Mar 2004)	First domain set (522/0/459)	Applied to predicted domains. Pfam (v. 10.0, trusted cutoff), BLAST (v. 2.2.4, $e=10^{-1}$), PSI-BLAST (v. 2.2.4, $h=10^{-2}$, snr dated 26 Feb 2004, $e=10^{-1}$).	Same as #3, but applied to predicted domains. No internal UGA codons allowed.	10

Table II: Coverage of *Mycoplasma pneumoniae* and *Mycoplasma genitalium* proteomes by structures solved prior establishment of the BSGC (Pre- BSGC row), and by all BSGC targets from the 6 rounds of target selection described in Table I. Parallelism indicates the average number of targets homologous to each *Mycoplasma pneumoniae* or *Mycoplasma genitalium* protein that is covered by at least one target. Residue coverage is calculated as a percentage of all residues, and as a percentage of the residues predicted to be HT-tractable and interesting (in parentheses).

Round	**Fine Coverage of *M. pneumoniae***			**Coarse Coverage of *M. pneumoniae***		
	Proteins (687 total)	Residues, % (239,722 total)	Parallelism	Proteins (687 total)	Residues, % (239,722 total)	Parallelism
Pre-BSGC	142 (20.7%)	16.5 (20.5)	n/a	297 (43.2%)	35.2 (43.1)	n/a
1	272 (39.6%)	26.6 (32.7)	1.4	424 (61.7%)	47.2 (56.8)	1.7
2	311 (45.3%)	30.7 (37.5)	3.2	467 (68.0%)	52.5 (62.6)	3.4
3	340 (49.5%)	34.0 (41.3)	3.4	493 (71.8%)	56.1 (66.7)	3.6
4	356 (51.8%)	35.7 (43.4)	3.5	495 (72.1%)	56.6 (67.3)	4.2
5	399 (58.1%)	45.0 (54.1)	3.6	532 (77.4%)	64.8 (76.0)	4.3
6	438 (63.8%)	48.2 (57.9)	8.6	550 (80.1%)	67.3 (78.7)	9.9
	Fine Coverage of *M. genitalium*			**Coarse Coverage of *M. genitalium***		
	Proteins (486 total)	Residues, % (175,930 total)	Parallelism	Proteins (486 total)	Residues, % (175,930 total)	Parallelism
Pre-BSGC	137 (28.2%)	21.5 (26.6)	n/a	262 (53.9%)	43.2 (52.4)	n/a
1	196 (40.3%)	29.4 (36.1)	1.6	311 (64.0%)	51.8 (62.0)	1.9
2	215 (44.2%)	31.9 (39.1)	1.9	328 (67.5%)	54.6 (65.1)	2.2
3	226 (46.5%)	33.8 (41.4)	1.9	340 (70.0%)	56.7 (67.6)	2.3
4	231 (47.5%)	34.5 (42.2)	2.2	341 (70.2%)	57.0 (68.0)	2.9
5	244 (50.2%)	38.4 (46.8)	2.2	354 (72.8%)	60.8 (72.1)	2.9
6	283 (58.2%)	42.3 (51.5)	4.6	371 (76.4%)	64.0 (75.9)	5.7

Table III: Coverage of *Mycoplasma pneumoniae* and *Mycoplasma genitalium* proteomes by structures solved prior to establishment of the BSGC (Pre- BSGC row), a cumulative total of structures solved at the BSGC and all structures solved prior to its establishment (+BSGC row), all structures solved outside the BSGC, including those solved prior to the establishment of the BSGC (Non- BSGC row), and by all current structures (Current). A relative timeline of these four groups, and a histogram illustrating the coarse coverage statistics, are shown in Figure 2. The Structures column indicates the number of entries from the knownstr database (i.e., PDB chains and structural genomics targets) that contributed to coverage in each row. The latter database includes some redundant entries; e.g., a PDB entry, a PDB "on-hold" sequence, and a structural genomics target might all refer to the same protein. Residue coverage is calculated as a percentage of all residues, and as a percentage of the residues predicted to be HT-tractable and interesting (in parentheses).

Set	**Fine Coverage of *M. pneumoniae***			**Coarse Coverage of *M. pneumoniae***		
	Structures	Proteins (687 total)	Residues, % (239,722 total)	Structures	Proteins (687 total)	Residues, % (239,722 total)
Pre-BSGC	1453	142 (20.7%)	16.5 (20.5)	3270	297 (43.2%)	35.2 (43.1)
+ BSGC	1569	162 (23.6%)	18.0 (22.4)	3452	326 (47.5%)	38.0 (46.5)
Non-BSGC	4285	240 (34.9%)	28.7 (35.4)	9816	380 (55.3%)	48.9 (58.3)
Current	4371	254 (37.0%)	29.8 (36.7)	9972	391 (56.9%)	49.8 (59.4)
	Fine Coverage of *M. genitalium*			**Coarse Coverage of *M. genitalium***		
	Structures	Proteins (486 total)	Residues, % (175,930 total)	Structures	Proteins (486 total)	Residues, % (175,930 total)
Pre-BSGC	1305	137 (28.2%)	21.5 (26.6)	2945	262 (53.9%)	43.2 (52.4)
+ BSGC	1405	154 (31.7%)	23.5 (29.0)	3124	287 (59.1%)	46.6 (56.4)
Non-BSGC	3976	233 (47.9%)	37.4 (45.7)	8970	339 (69.8%)	60.4 (70.7)
Current	4052	243 (50.0%)	38.6 (47.1)	9123	348 (71.6%)	61.5 (72.0)

Table IV: Pfam-A families corresponding to BSGC targets, for which the BSGC solved the first or only structures of proteins in the family. The PDB ID and date of PDB deposition are also shown. Some structures solved prior to the formal establishment of the BSGC which were selected as BSGC targets in round 1 are included; these structures have PDB deposition dates prior to 1 September 2000.

Families solved only at the BSGC				
Family Size	Accession	Family description	PDB	Date
208	PF01895	PhoU family	1SUM	26 Mar 2004
148	PF01513	ATP-NAD kinase	1SUW	26 Mar 2004
143	PF01515	Phosphate acetyl/butaryl transferase	1R5J	10 Oct 2003
92	PF02130	Uncharacterized protein family UPF0054	1OZ9	8 Apr 2003
91	PF02381	Domain of unknown function UPF0040	1N0E	13 Oct 2002
86	PF05175	Methyltransferase small domain	1DUS	18 Jan 2000
73	PF04079	Putative transcriptional regulators (Ypuh-like)	1T6S	7 May 2004
68	PF02635	DsrE/DsrF-like family	1JX7	5 Sep 2001
31	PF04327	Protein of unknown function (DUF464)	1S12	5 Jan 2004
26	PF04297	Putative HTH protein, YlxM/p13-like	1S70	29 Jan 2004
Families solved first at the BSGC, but later solved elsewhere				
Family Size	Accession	Family description	PDB	Date
617	PF00011	Hsp20/alpha crystallin family	1SHS	30 Jul 1998
551	PF00467	KOW motif	1EIF	29 Jul 1998
540	PF00582	Universal stress protein family	1MJH	4 Nov 1998
387	PF01965	DJ-1/PfpI family	1G2I	19 Oct 2000
150	PF02566	OsmC-like protein	1LQL	10 May 2002
141	PF01351	Ribonuclease HII	1EKE	7 Mar 2000
110	PF01812	5-formyltetrahydrofolate cyclo-ligase family	1SBQ	10 Feb 2004
109	PF01709	Domain of unknown function DUF28	1LFP	11 Apr 2002
105	PF01687	Riboflavin kinase / FAD synthetase	1MRZ	19 Sep 2002
104	PF01725	Ham1 family	2MJP	27 Jan 1999
99	PF02645	Uncharacterized protein, DegV family	1MGP	15 Aug 2002
88	PF01746	tRNA (Guanine-1)-methyltransferase	1OY5	3 Apr 2003
68	PF01287	Eukaryotic initiation factor 5A hypusine, DNA-binding OB fold	1EIF	29 Jul 1998
53	PF01269	Fibrillarin	1FBN	25 Apr 1999

Table V: Impact of BSGC structures on coverage of other organisms. Table IV lists 24 Pfam-A families for which the BSGC solved the first or only structures of members of the family; this group of families is referred to as Pfam-BSGC. Representation of those families in other proteomes, as well as Swiss-Prot (SP) and TrEMBL, is shown here.

Proteome / Set	**Total # of proteins in set**	**Proteins covered by Pfam-BSGC**	**Total # of interesting and HT-tractable residues**	**Residues covered by Pfam-BSGC**
A. thaliana	26,209	100 (0.4%)	9,613,448	13,733 (0.1%)
C. elegans	22,602	37 (0.2%)	7,709,635	4,104 (0.1%)
D. melanogaster	15,908	36 (0.2%)	6,848,099	5,495 (0.1%)
E. coli	4,357	36 (0.8%)	1,101,407	5,898 (0.5%)
H. sapiens	34,560	43 (0.1%)	12,502,002	5,003 (<0.1%)
M. jannaschii	1,777	20 (1.1%)	410,871	2495 (0.6%)
M. tuberculosis	3,877	33 (0.9%)	1,050,708	5,649 (0.5%)
M. musculus	38,795	66 (0.2%)	13,397,269	7,013 (0.1%)
R. norvegicus	27,479	40 (0.1%)	8,985,290	3,962 (<0.1%)
Swiss-Prot	127,046	1,122 (0.9%)	38,898,937	162,049 (0.4%)
SP+ TrEMBL	984,936	3,737 (0.4%)	249,695,988	532,320 (0.2%)

Table VI: Predicted biochemical and cellular roles of BSGC targets and ORFs from *M. pneumoniae* and *M. genitalium*. The first column shows the TIGR major role categories. The second column shows the total number of targets annotated in each role, along with the number solved and the number of currently active targets remaining. The last two columns show annotations of *Mycoplasma* proteomes: the first number in each column is the total number of proteins in the proteome in that role, the second is the number with some structural coverage at the "coarse" level, and the third is the number of proteins with "fine" structural coverage.

TIGR Role	**Targets / # solved / # active**	**Proteomes: # of proteins (total / coarse / fine)**	
		M. pneumoniae	*M. genitalium*
Amino acid biosynthesis	1/1/0	1/1/1	0/0/0
Biosynthesis of cofactors, prosthetic groups, and carriers	10/4/5	7/7/5	5/5/3
Cell envelope	60/0/57	50/15/0	28/7/1
Cellular processes	16/0/10	6/6/3	6/6/4
Central intermediary metabolism	11/2/1	8/8/7	7/7/7
DNA metabolism	121/1/119	36/29/17	28/26/18
Energy metabolism	28/2/9	37/37/27	32/32/25
Fatty acid and phospholipid metabolism	11/0/3	9/7/1	8/6/1
Protein fate	24/2/18	22/21/16	20/19/14
Protein synthesis	69/4/51	77/74/70	89/86/75
Purines, pyrimidines, nucleosides, and nucleotides	14/1/7	20/20/17	17/17/15
Regulatory functions	33/3/23	5/5/4	5/5/3
Signal transduction	0/0/0	0/0/0	0/0/0
Transcription	24/5/14	11/10/9	13/12/11
Transport and binding proteins	18/0/14	35/24/20	33/23/21
Hypothetical proteins	201/15/120	88/34/14	160/75/26
Unclassified function	225/4/206	162/51/18	1/0/0
Unknown function	58/1/43	12/11/7	12/12/11
* No annotation	24/1/21	101/31/18	22/10/8

Chandonia et al, Figure 1

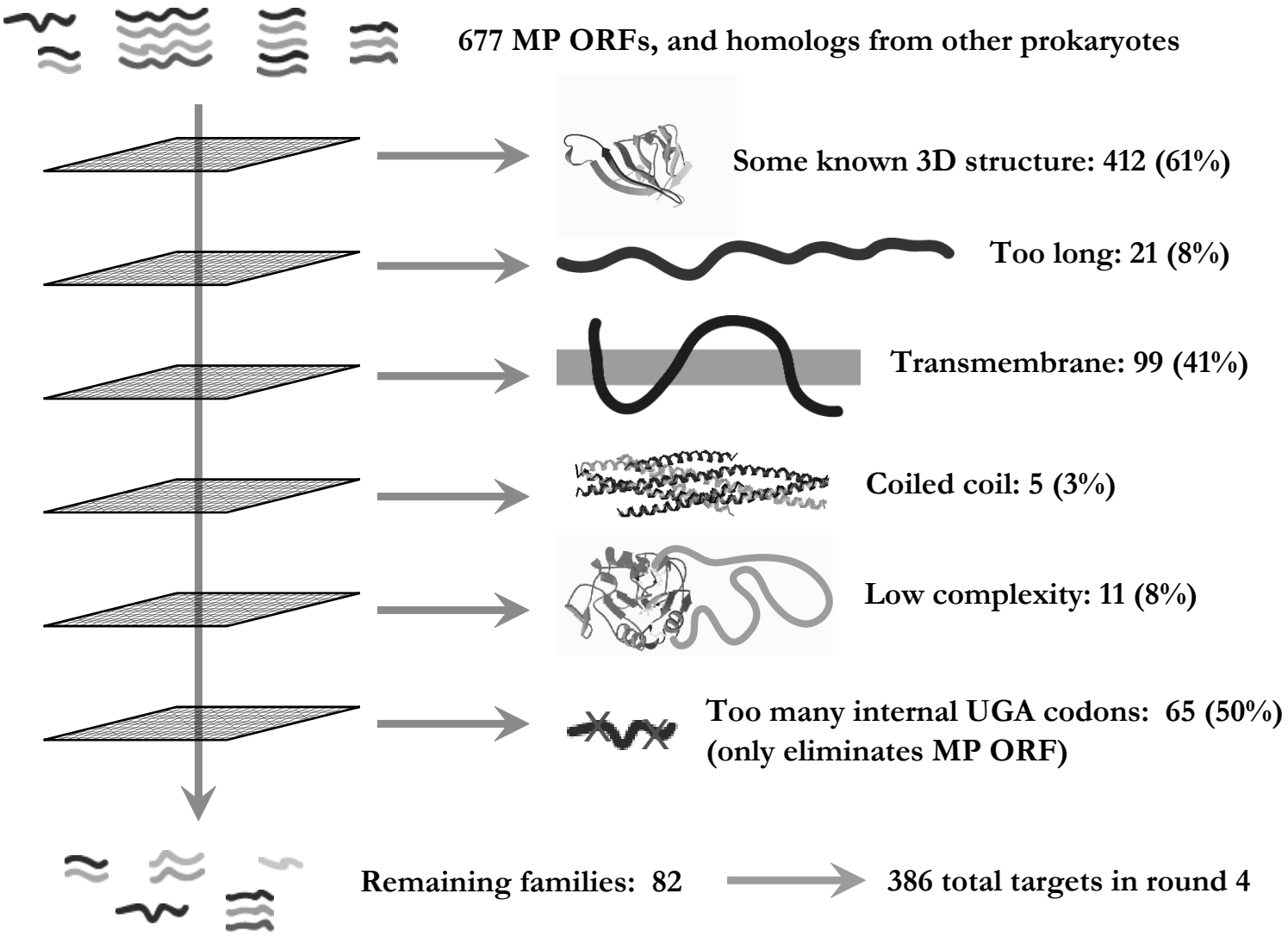


Figure 2: Coarse Coverage of *Mycoplasma* proteomes

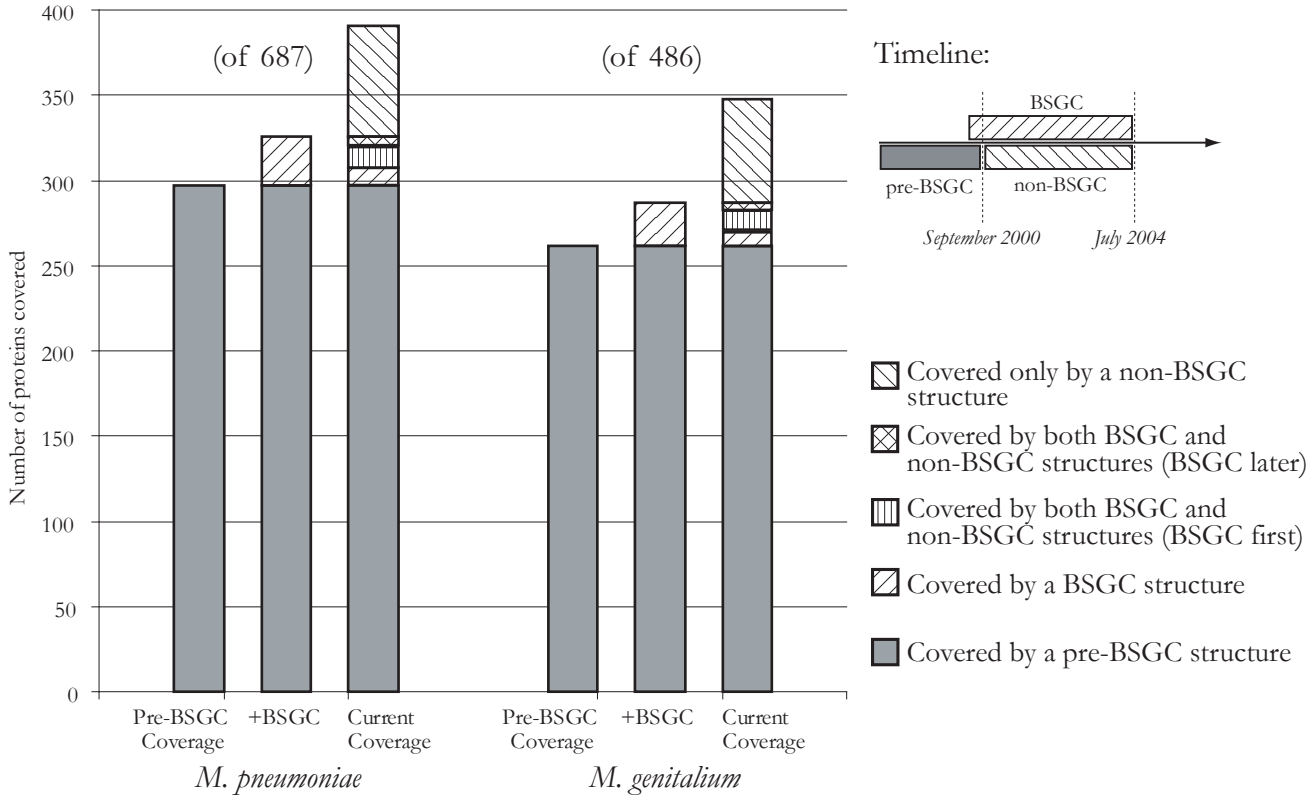


Figure 3: Percentage and number of BSGC targets stopped

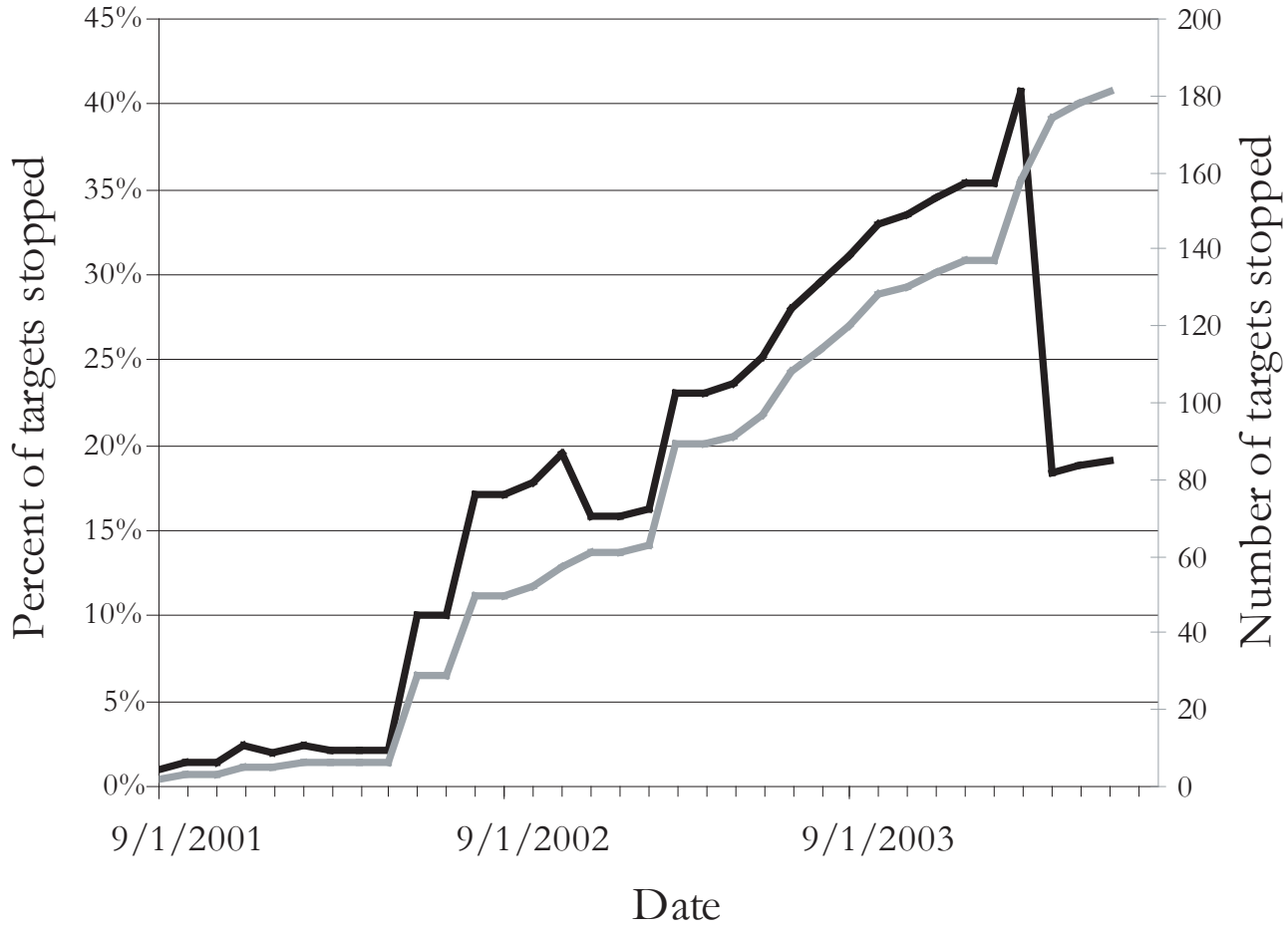
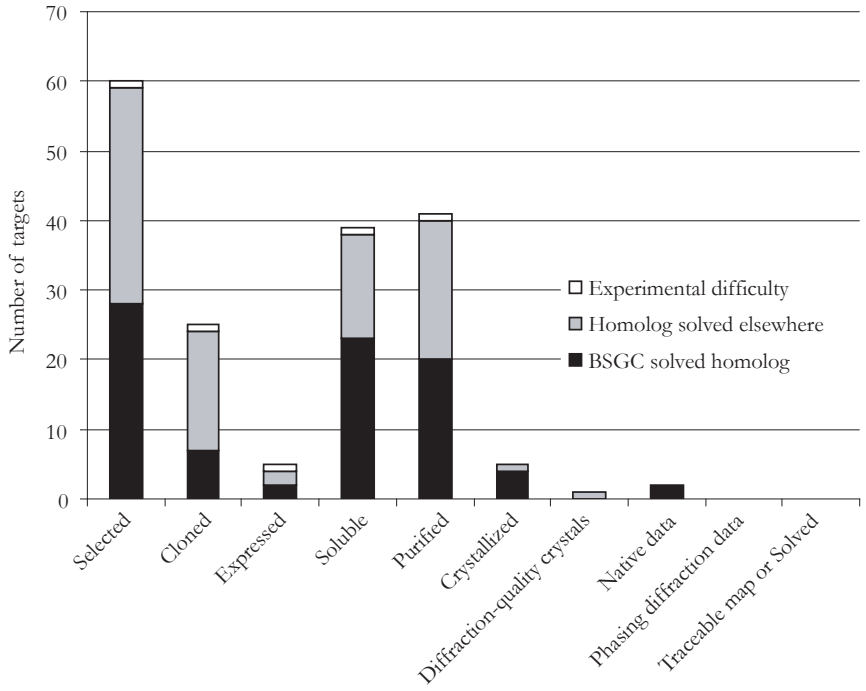


Figure 4: Stage and reason why targets were deselected



Target Selection and Deselection at the Berkeley Structural Genomics Center

Supplementary Information

John-Marc Chandonia[1], Sung-Hou Kim[1,2], and Steven E. Brenner[1,3]

Address for correspondence:
Steven E. Brenner
Department of Plant and Microbial Biology
461A Koshland Hall
University of California
Berkeley, CA 94720-3102
email: brenner@compbio.berkeley.edu
fax: (413) 280-7813

Affiliations:
1 - Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
2 - Department of Chemistry, University of California, Berkeley, CA 94720, USA
3 - Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

Target Organisms

The BSGC has selected targets from the prokaryotes shown in Table S1. The current list may be found on our website, http://www.strgen.org/

Experimental Attrition Rates

Table S2 shows the experimental status of BSGC targets as of 13 July 2004. As shown in Figure 4, only five of the targets deselected prior to 1 June 2004 were stopped due to experimental difficulties alone; most were stopped due to the BSGC or another group solving a homolog of the target. Therefore, the "In Progress" column of Table S2 emphasizes the current stages of the 649 targets that have were selected but not deselected.

As the data in Table S2 represents a snapshot in time only 4 months after the selection of the majority of the targets, the distribution of stages of these targets is too preliminary to be informative. However, the rightmost column in Table S2 contains a subset of data on the targets from the 3 automated rounds (2-4) of target selection, which were chosen between 28 August 2001 and 7 Nov 2002, and have been studied for at least 20 months. Of these 227 targets, 19 (8%) have been solved, and approximately half (115/227, or 51%) are still active. For these targets, the major experimental bottleneck appears to be crystallization. Of the 115 active targets, 43 were purified but not successfully crystallized, and 10 more have not achieved diffraction quality crystals. Overall, 25 of 78 purified proteins (32%) yielded sufficiently good crystals for diffraction. The next most significant bottleneck was purification of soluble proteins: of 102 soluble proteins, 78 (76%) were successfully purified.

A more recent analysis of the domain targets (round 6), was performed using results from 10 Feb 2005. This data, in Table S3, is instructive for comparing experimental

bottlenecks of domain and full-length targets. In the 7 months between the snapshots of Table S2 and S3, additional structures were solved for the full-length targets selected in rounds 2-4, and 30 more targets in this group were stopped due to homologs being solved at the BSGC or elsewhere. However, the major experimental bottlenecks for this group appear qualitatively similar to those indicated by the results from 13 July 2004. In contrast, although most of the domain targets were successfully cloned by 10 Feb 2005, expression of soluble protein has been a significant bottleneck for these targets. About one third (121 of 361) of the domain targets cloned were not successfully expressed, and over half of the expressed clones (127 of 240) were not successfully solublized. We expect that many of these cases are due to incorrectly predicted domain boundaries or domains that are unable to fold independently.

These results are preliminary and must be treated with caution, as there are indications that some of the recent experimental bottlenecks in the domain target set may be the result of using a new fusion tag. Direct comparison with other centers is difficult, as there are differing interpretations of the standards for targets reaching most of the experimental stages. However, we plan to perform a more complete analysis of bottlenecks, including comparison to other structural genomics centers, after more experimental work has progressed on the domain target set.

Table S1: Organisms from which BSGC targets were chosen.

Organism (Strain) Name	**Number of Targets**
Aeropyrum pernix	8
Allochromatium vinosum	3
Aquifex aeolicus	39
Archaeoglobus fulgidus	24
Bacillus halodurans	37
Bacillus subtilis	35
Borrelia burgdorferi	1
Campylobacter jejuni	3
Chlorobium tepidum TLS	16
Clostridium acetobutylicum	6
Deinococcus radiodurans	10
Escherichia coli K12	23
Escherichia coli O157:H7	1
Haemophilus influenzae Rd	5
Halobacterium sp. NRC-1	27
Helicobacter pylori 26695	5
Helicobacter pylori J99	2
Methanococcus jannaschii	57
Methanothermobacter thermautotrophicus	23
Mycoplasma genitalium	97
Mycoplasma pneumoniae	319
Neisseria meningitidis MC58	4
Neisseria meningitidis Z2491	3
Nostoc sp. PCC 7120	5
Pseudomonas aeruginosa	7
Pyrococcus furiosus DSM 3638	8
Pyrococcus horikoshii	11
Salmonella typhimurium LT2	3
Staphylococcus aureus subsp. *aureus* Mu50	9
Streptococcus agalactiae	7
Streptococcus pneumoniae R6	6
Streptococcus pneumoniae TIGR4	4
Streptococcus pyogenes	13
Streptomyces coelicolor A3(2)	1
Sulfolobus solfataricus	7
Thermoplasma acidophilum	10
Thermoplasma volcanium	9
Thermotoga maritima	72
Ureaplasma urealyticum	14
Vibrio cholerae	6
Xanthomonas campestris pv. *campestris* str. ATCC 33913	5
Xylella fastidiosa	1
Xylella fastidiosa 9a5c	2

Table S2: Current experimental stages of active and solved BSGC targets, as of 13 July 2004. The "Number of targets" column shows the total number of targets that have reached each experimental stage. In the "In Progress" columns, the first 8 rows show the number of targets at each stage that have not been deselected, and have not progressed to a subsequent stage. The current statistics may be found online at http://www.strgen.org/

Experimental Stage	**Number of Targets, All rounds**	**Targets in Progress, All rounds**	**Targets in Progress, Rounds 2-4 only**
Selected	945	231	12
Cloned	640	89	16
Expressed	499	119	4
Soluble	366	117	24
Purified	194	66	43
Crystallized	84	19	10
Diffraction quality crystals	60	7	5
Traceable map	49	1	1
Crystal structure	48 targets / 66 structures	48 targets / 66 structures	18 targets / 19 structures
NMR structure	3 targets / 3 structures	3 targets / 3 structures	1 target / 1 structure
Deselected	245	245	93

Table S3: Current experimental stages of active and solved BSGC targets, from target selection rounds described in this report, as of 10 Feb 2005. The "Number of targets" column shows the total number of targets that have reached each experimental stage. In the "In Progress" columns, the first 8 rows show the number of targets at each stage that have not been deselected, and have not progressed to a subsequent stage. The current statistics may be found online at http://www.strgen.org/

Experimental Stage	**Number of Targets, All rounds**	**Targets in Progress, Rounds 2-4 only**	**Targets in Progress, Round 6 only**
Selected	952	11	20
Cloned	820	9	121
Expressed	597	2	127
Soluble	420	14	78
Purified	239	35	30
Crystallized	94	9	2
Diffraction quality crystals	69	2	2
Traceable map	55	0	0
Crystal structure	55 targets / 81 structures	21 targets / 30 structures	1 target / 1 structure
NMR structure	3 targets / 3 structures	1 target / 1 structure	0
Deselected	356	123	141